# Introduction to Environmental Statistics

Professor Jessica Utts
University of California, Irvine
jutts@uci.edu

## A Few Sources for Data Examples Used

1. Statistical Methods in Water Resources by D.R. Helsel and R.M. Hirsch (H&H)
2. Statistical Methods and Pitfalls in Environmental Data Analysis, Yue Rong, *Environmental Forensics*, Vol 1, 2000, pgs 213-220. (Rong)
3. Data provided by Steve Saiz (Saiz)
4. Introduction to Probability and Statistics, Mendenhall and Beaver (M&B)
5. Occasionally, my own data because I understand it!



## The Big Picture

Why use statistical methods?

- Describe features of a data set
- Make inferences about a population using sample data
  - Estimate parameters (such as means) with a certain level of confidence
  - Test hypotheses about specific parameters or relationships
- Predict future values based on past data



## Populations and Samples

Population data:

- Measurements available on all "units" of interest.
  - Example: Annual peak discharge for Saddle River, NJ from 1925 to 1989. (H&H)
  - Can be considered as *population data* if only those years are of interest.
  - Can be considered as sample data and used for inference about *all* possible years.



## Sample Data

Sample data used for two purposes:

- Describing that sample only
- Making inferences to a population
- Ideal is a "random sample" but almost impossible to get. Instead:

**Fundamental Rule for Using Data for Inference:** Available data can be used to make inferences about a much larger group *if the data can be considered to be representative with regard to the question(s) of interest.*



## Examples of sample data

- Nickel effluent data, City of San Francisco (data from Saiz)
  - Grab samples from 1999 to 2002
  - Representative of a larger population of nickel concentration data; what population?
- Groundwater monitoring data for benzene concentrations for 16 quarters, 1996 to 1999 (data from Rong)
  - Representative of a larger population?

## Independent vs Paired Samples

For comparing two situations, data can be collected as independent samples or as "matched pairs." Examples:

- Independent samples:
  - Compare wells upgradient and downgradient from a toxic waste site for a certain chemical.
- Matched pairs (H&H):
  - Compare atrazine concentrations before (June) and after (Sept) application season in 24 shallow groundwater sites (same sites both times).



## Class Input and Discussion

Share examples of data you have collected and/or dealt with in your job:

- How were the data collected?
- What was the "question of interest?"
- Paired data or independent samples?
- Population or sample?
- If sample, what larger population do they represent?



## Types of Data

There are various ways to classify data, but probably not too useful for your data:

- Nominal:
  - "Name" only, also called *categorical data*
  - Example: Classify wells by land use in area (residential, agricultural, industrial, etc)
- Ordinal:
  - Ordered, but numbers may not have much meaning.
  - Example (H&H): 0 = concentrations below reporting limit, 1 = above rl but below a health standard, 2 = above health standard.



## Types of Data, Continued

- Interval:
  - Numbers have meaning, but ratios do not.
  - There is no absolute 0 ("none").
  - Example: Temperature
- Ratio:
  - There is an absolute 0
  - Ratios have meaning.
  - Example: Nickel concentration (it makes sense to talk about a sample having twice the concentration of another sample).



## Categorical vs Quantitative Data

- Categorical data
  - Nominal, and sometimes ordinal
  - For a single variable, summaries include frequencies and proportions only
- Quantitative data
  - Interval, ratio, sometimes ordinal
  - Summarize with numerical summaries
- Multiple variables (same or different types). For instance, categorize wells by land use, and compare a quantitative measure across uses.



## Types of Statistical Procedures

Graphical summaries
- Provide visual information about the data set
- Provide guidance on what statistical methods are appropriate

Numerical summaries
- Provide information about specific features of the data set

Inference (parametric, nonparametric)
- Infer things about a population, often to answer a yes/no question

# Summary Features of Quantitative Data

1. Location (Center, Average)
2. Spread (Variability)
3. Shape (Normal, skewed, etc)
4. Outliers (Unusual values)

We use pictures *and* numerical information to examine these.



## Questions about quantitative variables:

***One Quantitative Variable***

***Question 1:*** What interesting summary measures, like the average or the range of values, can help us understand the data?

***Example:*** What is the average nickel concentration in the SF effluent data, and how much variability is there?

***Question 2:*** Are there individual data values that provide interesting information because they are unique or stand out in some way (outliers)?

***Example:*** (M&B) Data on mercury concentration in livers of 28 dolphins were all over 100 micrograms/gram except 4 of them, which were all under 10. Explanation: 4 dolphins under 3 years old, others all more than 8 years old.



## One Categorical, One Quantitative Variable (Comparing across categories)

***Question 1:*** Are the quantitative measurements similar across categories of the categorical variable?

***Example:*** (H&H) Do wells upgradient and downgradient of a hazardous waste site have the same average concentration of a toxic compound?

***Question 2:*** When the categories have a natural ordering (an ordinal variable), does the quantitative variable increase or decrease, on average, in that same order?

***Example:*** Do low, medium and high flow areas of a stream have an increasing (or decreasing) average amount of a certain type of vegetation?



## Pictures for Quantitative Data

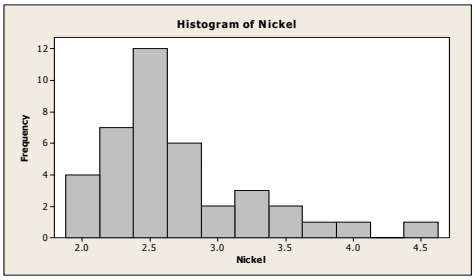
- **Look at** shape, outliers, center (location), spread, gaps, any other interesting features.

**Four common types of pictures:**

- **Histograms**: similar to bar graphs, used for *any number* of data values.
- **Stem-and-leaf plots** and **dotplots**: present *all individual values*, useful for *small to moderate* sized data sets.
- **Boxplot** or **box-and-whisker plot**: useful *summary* for *comparing* two or more groups.



## Histogram: SF Nickel Effluent Data (Saiz)



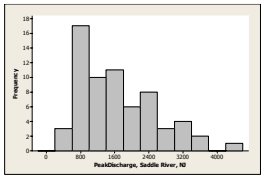

- Values are "centered" at about 2.6 or 2.7 (µg/L)
- Shape is "skewed to the right" (more on this later)
- Values range from about 1.9 to 4.5

## Notes about histograms

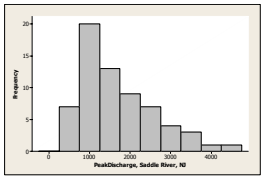
- Intervals are equally spaced. Example: Each interval has width 0.25.
- One goal is to assess shape. Between 6 and 15 intervals is a good number (may need to use more if there are gaps and/or outliers). 11 in nickel example.
- Some authors suggest using smallest $k$ with $2^k \geq n$, but not good for small n. Ex: n =39, so would use only $k = 6$.
- Decide where to put values that are on the boundary. For instance, would 2 go in the interval from 0 to 2, or from 2 to 4? Need to be consistent. (Not relevant in this example.)
- Can use *frequencies* (counts) or *relative frequencies* (proportions) as vertical axis. Example uses frequencies.

Number of "bins" can change picture
Ex: Peak discharge, Saddle River, NJ (H&H)

12 intervals




10 intervals




Even a small change in number of intervals made a difference.

# Creating a Dotplot

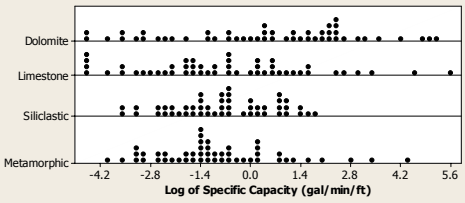
**These can be useful for comparing groups**

- Ideally, number line represents all possible values and there is one dot per observation. Not always possible. If dots represent multiple observations, footnote should explain that.
- As with histogram, divide horizontal axis into equal intervals, then put dots on it for each individual in each interval.
- Example (next slide): Compare ln(specific capacity) for wells in Appalachians of Pennsylvania, 4 rock types. [*ln(x) = natural log of x*]



## Dotplots of Ln(specific capacity), H&H

Each dot represents one observation. Note different ranges, and possibly different centers.




# Creating a Stemplot (stem and leaf plot) Ex: SF nickel effluent data

2.8, 3.0, 3.3, 2.5, 2.3, 2.4, 2.7, 2.8, 2.6, 3.9, 3.5, 2.5, 3.7, **4.4**, 2.3, 2.6, 2.5, 2.2, 2.6, 3.2, 3.0, **1.9**, 2.3, 2.3, 3.5, 2.4, 2.2, 2.4, 2.4, 2.2, 2.0, 2.5, 2.8, 2.7, 2.8, 2.1, 2.6, 3.3, 2.1

## Step 1: Create the Stem

Divide range of data into equal units to be used on **stem**. Have 6 to 15 stem values, representing *equally spaced* intervals. Here, we could use 2 or 5 for each digit from 1 to 4.

1|
2|
2|
3|
3|
4|

**Example:** Each of the 6 stem values represents a possible range of 0.5 First one represents 1.5 to 1.9, then 2.0 to 2.4, then 2.5 to 2.9, and so on, up to 4.0 to 4.4.

# Creating a Stemplot

## Step 2: Attach the Leaves

Attach a **leaf** to represent each data point. Next digit in number used as leaf; drop any remaining digits.

**Example:** First 5 values are 2.8, 3.0, 3.3, 2.5, 2.3. The numbers after the decimal point are the "leaves"

1|
2| 3
2| 8 5
3| 0 3
3|
4|

**Optional Step:** order leaves on each branch.



# Further Details for Creating Stemplots

Reusing digits two or five times. Goal: assess shape.

**Stemplot A:**
1|9
2|01122233334444
2|55556666778888
3|00233
3|5579
4|4

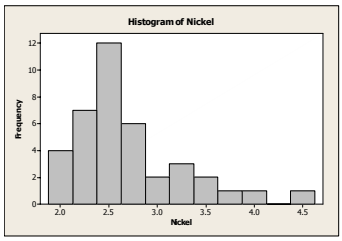
EX: 1|9 = 1.9

Two times:
1st stem = leaves 0 to 4
2nd stem = leaves 5 to 9

**Stemplot B:**
1|9
2|011
2|2223333
2|44445555
2|666677
2|8888
3|00
3|233
3|55
3|7
3|9
4|
4|
4|4

Five times:
1st stem = leaves 0 and 1
2nd stem =leaves 2 and 3, etc.

## Nickel Example: Shape




This shape is called "skewed to the right."

**Stemplot B:**

1|9
2|011
2|2223333
2|44445555
2|666677
2|8888
3|00
3|233
3|55
3|7
3|9
4|
4|
4|4

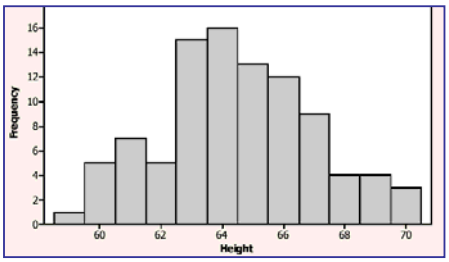
## Describing Shape

- **Symmetric**, **bell-shaped**
- **Symmetric**, ***not*** **bell-shaped**
- **Bimodal:** Two prominent "peaks" (modes)
- **Skewed Right**: On number line, values clumped at left end and *extend* to the *right* (Very common in your data sets.)
- **Skewed Left**: On number line, values clumped at right end and *extend* to the *left* (Ex: Age at death from heart attack.)



## Bell-shaped example: Heights of 94 females

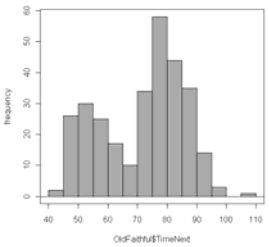
**Heights** of 94 female college students. Bell-shaped, centered around 64 inches with no outliers.






## Bimodal Example: Old Faithful Geyser, time *between* eruptions, histogram from R Commander

**Times between eruptions** of the Old Faithful geyser, shape is **bimodal**. Two clusters, one around 50 min., other around 80 min.




**Source: Hand et al., 1994**



## Boxplots, based on "Five Number Summary":

***The five-number summary display***

| | Median | |
|---|---|---|
| Lower Quartile | | Upper Quartile |
| Lowest | | Highest |

- **Lowest** = Minimum
- **Highest** = Maximum
- **Median** = number such that half of the values are at or above it and half are at or below it (middle value or average of two middle numbers in ordered list).
- **Quartiles** = medians of the two halves.



## Boxplots

**Visual picture of the five-number summary**

**Example: How much do statistics students sleep?**

190 statistics students asked how many hours they slept the night before (a Tuesday night).

***Five-number summary for number of hours of sleep (details of how to find these a little later)***

| | | |
|---|---|---|
| | 7 | |
| 6 | | 8 |
| 3 | | 16 |

Two students reported 16 hours; the max for the remaining 188 students was 12 hours.

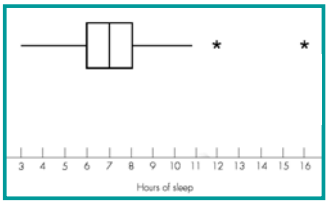
## Creating a Boxplot

1. Draw horizontal (or vertical) line, label it with values from lowest to highest in data.
2. Draw rectangle (box) with ends at quartiles.
3. Draw line in box at value of median.
4. Compute IQR = distance between quartiles.
5. Compute 1.5(IQR); *outlier* is any value more than this distance from closest quartile. Draw line (whisker) from each end of box extending to farthest data value that is not an outlier. (If no outlier, then to min and max.)
6. Draw asterisks to indicate the outliers.



## Creating a Boxplot for Sleep Hours

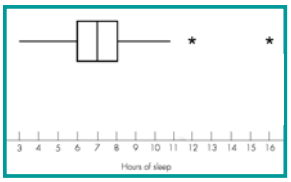
1. Draw horizontal line and label it from 3 to 16.
2. Draw rectangle (box) with ends at 6 and 8 (quartiles).
3. Draw line in box at median of 7.
4. Compute IQR (interquartile range) = 8 – 6 = 2.
5. Compute 1.5(IQR) = 1.5(2) = 3; outlier is any value below 6 – 3 = 3, or above 8 + 3 = 11.
6. Draw line from each end of box extending down to 3 and up to 11.
7. Draw asterisks at outliers of 12 and 16 hours.






## Interpreting Boxplots

- Divides the data into fourths.
- Easily identify outliers.
- Useful for comparing two or more groups.




**Outlier**: any value more than 1.5(IQR) beyond closest quartile.

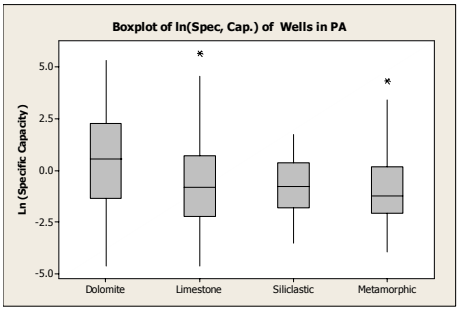
¼ of students slept between 3 and 6 hours

¼ slept between 6 and 7 hours

¼ slept between 7 and 8 hours

¼ slept between 8 and 16 hours

Sometimes boxplots are vertical instead of horizontal; also, useful for comparisons




## Outliers and How to Handle Them

**Outlier:** a data point that is not consistent with the bulk of the data.

- Look for them via graphs.
- Can have big influence on conclusions.
- Can cause complications in some statistical analyses.
- Cannot discard without justification.
- May indicate that the underlying population is skewed, rather than one unique outlier (especially with small samples)

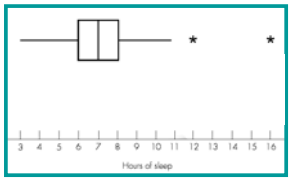


## Possible reasons for outliers and what to do about them:

1. *Outlier is legitimate data value and represents natural variability for the group and variable(s) measured.* Values may not be discarded. They provide important information about location and spread.
2. *Mistake made while taking measurement or entering it into computer.* If verified, should be discarded or corrected.
3. *Individual observation(s) in question belong(s) to a different group than bulk of individuals measured.* Values may be discarded if summary is desired and reported for the majority group only.

## Example: *Sleep hours*

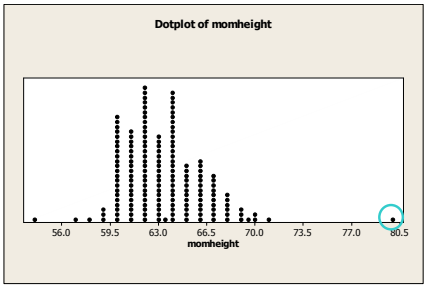
Two students were outliers in amount of sleep, but the values were not mistakes.




**Reason 1**: Natural variability, it is *not* okay to remove these values.



## Example: *Students gave mother's height*




Height of 80 inches = 6 ft 8 inches, almost surely an error!
**Reason #2**, investigate and try to find error; remove value.

## **Example:** Weights (in pounds) of 18 men on crew teams:

*Cambridge:* 188.5, 183.0, 194.5, 185.0, 214.0, 203.5, 186.0, 178.5, **109.0**

*Oxford:* 186.0, 184.5, 204.0, 184.5, 195.5, 202.5, 174.0, 183.0, **109.5**

**Note:** last weight in each list is unusually small. ???



## **Example:** Weights (in pounds) of 18 men on crew teams:

*Cambridge:* 188.5, 183.0, 194.5, 185.0, 214.0, 203.5, 186.0, 178.5, **109.0**

*Oxford:* 186.0, 184.5, 204.0, 184.5, 195.5, 202.5, 174.0, 183.0, **109.5**

**Note:** last weight in each list is unusually small. ???

They are the *coxswains* for their teams, while others are *rowers*.

**Reason 3**: different group, okay to remove if only interested in rowers.



## Numerical Summaries of Quantitative Data

**Notation for Raw Data:**

$n$ = number of individual observations in a data set
$x_1, x_2, x_3, \ldots, x_n$ represent individual raw data values

***Example:*** Nickel effluent data:
So $n = 39$, and

$x_1 = 2.8, \quad x_2 = 3.0, \quad x_3 = 3.3, \quad x_4 = 2.5$ etc....

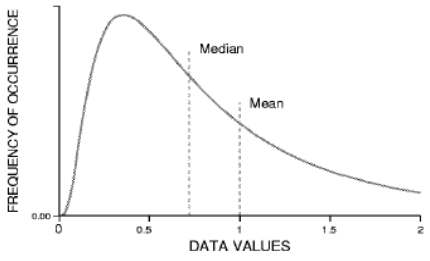


## Describing the "Location" of a Data Set

- **Mean:** the numerical average
- **Median:** the middle value (if $n$ odd) or the average of the middle two values ($n$ even)

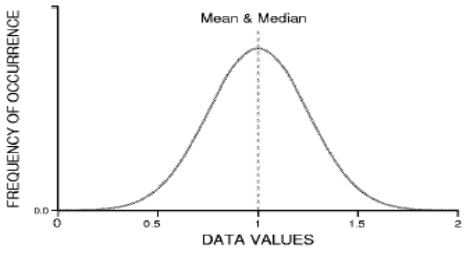
*Symmetric*: mean = median
*Skewed Left*: usually mean < median
*Skewed Right*: usually mean > median

## Pictures from Helsel &Hirsch: Data values skewed to the right






## Bell-shaped distribution






## Determining the Mean and Median

**The Mean** $\bar{x} = \dfrac{\sum x_i}{n}$

where $\sum x_i$ means "add together all the values"

**The Median**

If *n* is odd: *Median* = middle of ordered values.
Count $(n + 1)/2$ down from top of ordered list.

If *n* is even: *Median* = average of middle two ordered values. Average the values that are $(n/2)$ and $(n/2) + 1$ down from top of ordered list.



## The Mean, Median, and Mode

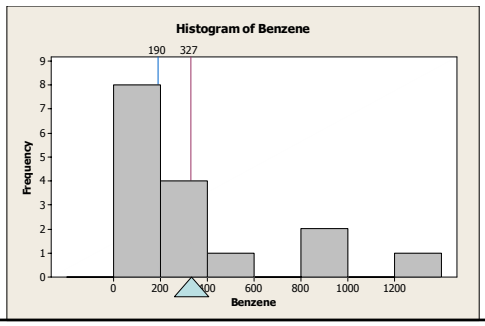
**Ordered Listing of 16 Benzene values (Rong)**

3.8, 35, 38, 55, 110, 120, 130, 180,
200, 230, 320, 340, 480, 810, 980, 1200

- **Mean (numerical average): 327.0**
- **Median: 190 (halfway between 180 and 200)**
- **Mode (most common value):** no single mode



## Median (190, half of area) vs Mean (327, balance point) Skewed to Right




## The Influence of Outliers on the Mean and Median

- **Larger influence on mean** than median.
- High outliers and data skewed to the *right* will increase the mean.
- Low outliers and data skewed to the *left* will decrease the mean.

**Ex:** Suppose ages at death of your eight great-grandparents are: 28, 40, 75, 78, 80, 80, 81, 82.

**Mean** age is 544/8 = 68 years old

**Median** age is (78 + 80)/2 = 79 years old

## Caution: Confusing *Normal with Average*

Common mistake to confuse "average" with "normal".

*Is woman 5 ft. 10 in. tall 5 inches taller than normal??*

**Example: How much hotter than normal is normal?**

> "October came in like a dragon Monday, hitting 101 degrees in Sacramento by late afternoon. That temperature tied the record high for Oct. 1 set in 1980 – and was 17 degrees *higher than normal for the date*. (Korber, 2001, italics added.)"

Article had thermometer showing "*normal* high" for the day was 84 degrees. High temperature for Oct. 1st is quite variable, from 70s to 90s. While 101 was a record high, it was not "17 degrees higher than normal" if "normal" includes the range of possibilities likely to occur on that date.



## Describing Spread (Variability): Range, Interquartile Range and Standard deviation

- **Range** = high value – low value
- **Interquartile Range (IQR)** = upper quartile – lower quartile = $Q_3 - Q_1$ (to be defined)
- **Standard Deviation** (most useful for bell-shaped data)



## Benzene Example, n = 16

3.8, 35, 38, 55,
[$Q_1$ = (55+110)/2 = 82.5]
110, 120, 130, 180,
[Median = 190]
200, 230, 320, 340,
[$Q_3$ = (340+480)/2 = 410]
480, 810, 980, 1200

| Five number summary | | |
|---|---|---|
| | 190 | |
| 82.5 | | 410 |
| 3.8 | | 1200 |

- ***Median*** = 190 has half of the values above, half below
- Two ***extremes*** describe spread over 100% of data
  ***Range*** = 1200 – 3.8 = 1196.2
- Two ***quartiles*** describe spread over middle 50% of data
  ***Interquartile Range*** = 410 – 82.5 = 327.5

## Finding Quartiles "by hand"

Split the ordered values at median:

- half at or below the median ("at" if ties)
- half at or above the median

$Q_1$ = **lower quartile**
= median of data values that are (at or) *below* the median

$Q_3$ = **upper quartile**
= median of data values that are (at or) *above* the median



## Hands-On Activity #1
## Data and details on activity sheet

- For the San Francisco effluent nickel data:
  - Find a 5-number summary
  - Draw a boxplot
- What can be concluded about shape from the boxplot?



## Results given in class

- Five number summary:
- Boxplot:
- Shape:

## Percentiles

The ***k*th percentile** is a number that has *k*% of the data values at or below it and (100 – *k*)% of the data values at or above it.

- Lower quartile: 25th percentile
- Median: 50th percentile
- Upper quartile: 75th percentile



## Describing Spread (Variability):

- **Range** = high value – low value
- **Interquartile Range (IQR)** = upper quartile – lower quartile = $Q_3 - Q_1$
- **Standard Deviation** – most useful for bell-shaped data



## Describing Spread with Standard Deviation

**Standard deviation** measures variability by summarizing how far individual data values are from the mean.

Think of the standard deviation as ***roughly the average distance values fall from the mean***.



## Describing Spread with Standard Deviation: A very simple example

| Numbers | Mean | Standard Deviation |
|---|---|---|
| 100,100,100,100,100 | 100 | 0 |
| 90, 90,100,110,110 | 100 | 10 |

Both sets have same mean of 100.

Set 1: all values are equal to the mean so there is *no variability* at all.

Set 2: one value equals the mean and other four values are 10 points away from the mean, so the *average distance away from the mean is about* 10.



## Calculating the Standard Deviation

Formula for the (***sample***) **standard deviation**:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

The value of $s^2$ is called the (*sample*) **variance.**
An equivalent formula, easier to compute, is:

$$s = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}}$$



## Calculating the Standard Deviation

**Example: 90, 90, 100, 110, 110**

**Step 1:** Calculate $\bar{x}$ , the sample mean. *Ex*: $\bar{x} = 100$

**Step 2:** For each observation, calculate the difference between the data value and the mean.
*Ex*: -10, -10, 0, 10, 10

**Step 3:** Square each difference in step 2.
*Ex:* 100, 100, 0, 100, 100

**Step 4:** Sum the squared differences in step 3, and then divide this sum by $n - 1$. Result = *variance* $s^2$
*Ex:* 400/(5 – 1) = 400/4 = 100

**Step 5:** Take the square root of the value in step 4.
*Ex: s = standard deviation* = $\sqrt{100} = 10$

## Population Standard Deviation

Data sets usually represent a sample from a larger population. If the data set includes measurements for an ***entire population***, the notations for the mean and standard deviation are different, and the formula for the standard deviation is also slightly different.
A **population mean** is represented by the Greek $\mu$ ("mu"), and the **population standard deviation** is represented by the Greek "sigma" (lower case)

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$



## Bell-shaped distributions

- Measurements that have a bell-shape are so common in nature that they are said to have a *normal distribution*.
- Knowing the mean and standard deviation *completely determines* where all of the values fall for a normal distribution, assuming an infinite population!
- In practice we don't have an infinite population (or sample) but if we have a large sample, we can get good approximations of where values fall.

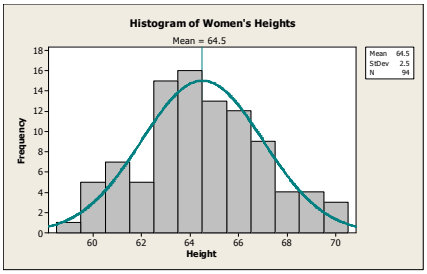


## Examples of bell-shaped data

- Women's heights
  - mean = 64.5 inches, s = 2.5 inches
- Men's heights
  - mean = 70 inches, s = 3 inches
- IQ scores
  - mean = 100, s = 15 (or for some tests, 16)



Women's heights, n = 94 students
Note approximate bell-shape of histogram
"Normal curve" with mean = 64.5, s = 2.5
superimposed over histogram




## Interpreting the Standard Deviation for Bell-Shaped Curves: The Empirical Rule

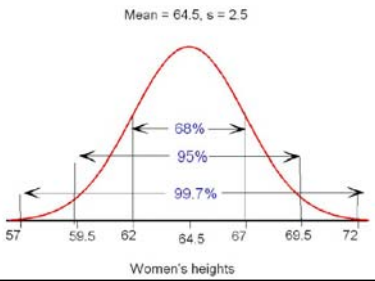
For any bell-shaped curve, approximately

- **68%** of the values fall within **1 standard deviation** of the mean in either direction
- **95%** of the values fall within **2 standard deviations** of the mean in either direction
- **99.7%** (almost all) of the values fall within **3 standard deviations** of the mean in either direction



## Ex: Population of women's heights

- 68% of heights are between 62 and 67 inches
- 95% of heights are between 59.5 and 69.5 inches
- 99.7% of heights are between 57 and 72 inches



## *Women's Heights: How well does the Empirical Rule work?*

Mean height for the 94 students was 64.5, and the standard deviation was 2.5 inches. Let's compare actual with ranges from Empirical Rule:

| Range of Values: | Empirical Rule | Actual number | Actual percent |
|---|---|---|---|
| Mean ± 1 s.d. | 68% in 62 to 67 | 70 | 70/94 = 74.5% |
| Mean ± 2 s.d. | 95% in 59.5 to 69.5 | 89 | 89/94 = 94.7% |
| Mean ± 3 s.d. | 99.7% in 57 to 72 | 94 | 94/94 = 100% |

## The Empirical Rule, the Standard Deviation, and the Range

- Empirical Rule tells us that the range from the minimum to the maximum data values equals about 4 to 6 standard deviations for data sets with an approximate bell shape.
- *For a large data set, you can get a rough idea of the value of the standard deviation by dividing the range by 6 (or 4 or 5 for a smaller dataset)*

$$s \approx \frac{Range}{6}$$



## Standardized *z*-Scores

**Standardized score** or ***z*-score**:

$$z = \frac{\text{Observed value} - \text{Mean}}{\text{Standard deviation}}$$

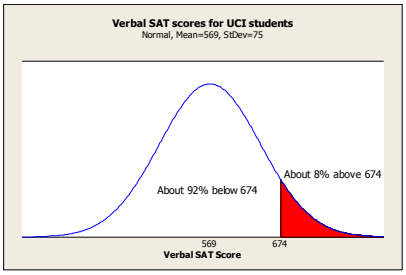
***Example:*** UCI Verbal SAT scores had mean = 569 and s = 75. Suppose someone had SAT = 674:

$$z = \frac{674 - 569}{75} = +1.40$$

Verbal SAT of 674 for UCI student is 1.40 standard deviations ***above*** the mean for UCI students.



Verbal SAT of 674 is 1.40 standard deviations above mean.
To find proportion above or below, use Excel or R Commander






## The Empirical Rule Restated for Standardized Scores (z-scores):

For bell-shaped data,

- About **68%** of the values have *z*-scores between –1 and +1.
- About **95%** of the values have *z*-scores between –2 and +2.
- About **99.7%** of the values have *z*-scores between –3 and +3.